Title

Ancient Origin and Recent Innovations of RNA Polymerase IV and V.

Journal

Molecular biology and evolution, 32(7)

ISSN

0737-4038

Authors

Huang, Yi
Kendall, Timmy
Forsythe, Evan S
et al.

Publication Date

2015-07-01

DOI

10.1093/molbev/msv060

Peer reviewed

Ancient Origin and Recent Innovations of RNA Polymerase IV and V

Yi Huang,[1] Timmy Kendall,[1] Evan S. Forsythe,[1] Ana Dorantes-Acosta,[2] Shaofang Li,[3] Juan Caballero-Pérez,[4] Xuemei Chen,[3] Mario Arteaga-Vázquez,[2] Mark A. Beilstein,[1] and Rebecca A. Mosher*[,1,5]

[1]The School of Plant Sciences, The University of Arizona

[2]Instituto de Biotecnología y Ecología Aplicada (INBIOTECA), Universidad Veracruzana, Veracruz, México

[3]Department of Botany and Plant Sciences, Institute of Integrative Genome Biology, University of California, Riverside

[4]Facultad de Ingeniería, Universidad Autónoma de Querétaro, Querétaro, México

[5]The Bio5 Institute, The University of Arizona

***Corresponding author:** E-mail: rmosher@email.arizona.edu.

Associate editor: Juliette de Meaux

Abstract

Small RNA-mediated chromatin modification is a conserved feature of eukaryotes. In flowering plants, the short interfering (si)RNAs that direct transcriptional silencing are abundant and subfunctionalization has led to specialized machinery responsible for synthesis and action of these small RNAs. In particular, plants possess polymerase (Pol) IV and Pol V, multi-subunit homologs of the canonical DNA-dependent RNA Pol II, as well as specialized members of the RNA-dependent RNA Polymerase (RDR), Dicer-like (DCL), and Argonaute (AGO) families. Together these enzymes are required for production and activity of Pol IV-dependent (p4-)siRNAs, which trigger RNA-directed DNA methylation (RdDM) at homologous sequences. p4-siRNAs accumulate highly in developing endosperm, a specialized tissue found only in flowering plants, and are rare in nonflowering plants, suggesting that the evolution of flowers might coincide with the emergence of specialized RdDM machinery. Through comprehensive identification of RdDM genes from species representing the breadth of the land plant phylogeny, we describe the ancient origin of Pol IV and Pol V, suggesting that a nearly complete and functional RdDM pathway could have existed in the earliest land plants. We also uncover innovations in these enzymes that are coincident with the emergence of seed plants and flowering plants, and recent duplications that might indicate additional subfunctionalization. Phylogenetic analysis reveals rapid evolution of Pol IV and Pol V subunits relative to their Pol II counterparts and suggests that duplicates were retained and subfunctionalized through Escape from Adaptive Conflict. Evolution within the carboxy-terminal domain of the Pol V largest subunit is particularly striking, where illegitimate recombination facilitated extreme sequence divergence.

Key words: **small RNA-directed DNA methylation, RNA Polymerase IV, RNA Polymerase V, Escape from Adaptive Conflict, Gene duplication, RNA Silencing.**

Article

Introduction

Eukaryotes encode three DNA-dependent RNA polymerases (Pol I, Pol II, and Pol III) for transcription of ribosomal, messenger, and transfer RNAs. Plants contain two additional polymerases, Pol IV and V, which are specialized for transcriptional gene silencing via RNA-directed DNA methylation (RdDM) (Herr et al. 2005; Kanno et al. 2005; Onodera et al. 2005; Pontier et al. 2005). Pol IV and V are related to Pol II and likely arose through subfunctionalization of silencing activities performed by Pol II in fungi and metazoans (Luo and Hall 2007; Ream et al. 2013). However, Pol IV and V might also have novel activities based within the C-terminal domain (CTD), which is unrelated to that of Pol II.

Pol IV initiates the synthesis of short interfering (si)RNAs from thousands of repetitive genomic loci, including all classes of transposons (Mosher et al. 2008). Pol IV physically associates with RNA-dependent RNA polymerase RDR2, which uses the Pol IV transcript as a template to generate double-stranded RNA (Haag et al. 2012). This double-stranded RNA is cleaved by the Dicer-like endonuclease DCL3, resulting in characteristic 24-nt siRNAs (Xie et al. 2004). Pol IV-dependent (p4-)siRNAs integrate into three ARGONAUTE (AGO) proteins (AGO4, AGO6, and AGO9) (Havecker et al. 2010), and the AGO/p4-siRNA complex is hypothesized to associate with specific genomic regions through Watson–Crick base pairing between the p4-siRNA and nascent, noncoding transcripts generated by Pol V (Wierzbicki et al. 2008, 2009). AGO4 physically associates with a WG/GW platform in the CTD of Pol V, presumably to aid AGO4 recruitment to chromatin (Li et al. 2006; El-Shami et al. 2007). AGO/p4-siRNA/Pol V complex assembles additional proteins to initiate DNA methylation and transcriptional silencing (Matzke and Mosher 2014).

In angiosperms (flowering plants), p4-siRNAs are the most abundant class in the small RNA transcriptome, comprising

up to 90% of the mass and 99% of the complexity in some tissues (Henderson et al. 2006; Kasschau et al. 2007; Zhang et al. 2007; Mosher et al. 2008). p4-siRNAs are abundant in all tissues, but are particularly prevalent in the developing endosperm, where they accumulate predominantly from matrigenic chromosomes (Mosher et al. 2009). This observation suggests that p4-siRNAs might mediate interactions or facilitate conflict between the matrigenic and patrigenic genomes in the endosperm (Mosher 2010).

Twenty-three nucleotide siRNAs capable of targeting DNA methylation to repetitive elements are present in moss (Cho et al. 2008), suggesting that small RNA-mediated transcriptional silencing is conserved throughout the land plant lineage. However, in moss these siRNAs are at low levels compared with 21 nt siRNAs (Cho et al. 2008), indicating that 24 nt siRNA expression is minimal or limited to a subset of tissues in nonflowering plants. Among sampled gymnosperms, there are conflicting reports of the presence of 24 nt siRNAs. Initially determined to be absent from conifers (Dolgosheina et al. 2008; Morin et al. 2008), recent publications identified 24 nt siRNAs in Chinese fir (*Cunninghamia lanceolata*) (Wan et al. 2012), Japanese larch (*Larix leptolepis*) (Zhang et al. 2013), and Norway spruce (*Picea abies*) (Nystedt et al. 2013). In these species, 24 nt siRNAs were found only in reproductive tissues or samples containing reproductive tissues. The limited production of 24 nt siRNAs in nonflowering plants indicates limited activity of Pol IV and Pol V in these species relative to angiosperms. Combined with the prevalence of p4-siRNAs in the endosperm (a tissue found only in angiosperms), this observation suggests that the evolution of an endosperm might coincide with changes in Pol IV/V function and/or structure.

Each RNA polymerase is a large holoenzyme complex composed of at least 12 subunits (Ream et al. 2009). Some subunits are shared by all five eukaryotic polymerases, while others are uniquely incorporated into a single polymerase and presumably grant functional specificity to the enzyme (Huang et al. 2009; Lahmy et al. 2009; Ream et al. 2009, 2013; Haag et al. 2014). Specific subunits of Pol IV and V (named NRPD and NRPE, respectively, for Nuclear RNA Polymerase D and E) arose from the duplication of Pol II (NRPB) subunits (Luo and Hall 2007; Tucker et al. 2010). Many subunits are shared by all three polymerases, but the largest/first subunits (NRPB1, NRPD1, and NRPE1) are unique to each polymerase (Ream et al. 2009; Haag et al. 2014). Additionally, Pol IV and V share second, fourth, fifth, and seventh subunits that are distinct from the Pol II versions, and there is evidence for continued duplication of subunits in specific lineages (Sidorenko et al. 2009; Stonaker et al. 2009; Tucker et al. 2010; Tan et al. 2012; Haag et al. 2014). The earliest duplication of an NRPB subunit was detected in Characeae, an algal lineage closely related to land plants, and additional NRPB to NRPD duplications occurred after the divergence of land plants from algae (Luo and Hall 2007). However, NRPD4/E4, NRPE1, and NRPE5 have been identified only in angiosperms (Luo and Hall 2007; Tucker et al. 2010). Additionally, phylogenomic analysis indicates that loss of NRPD2/E2 might be an important node for evolution of gymnosperms (Lee et al. 2011).

To better understand the composition of the Pol IV and Pol V holoenzymes in nonflowering plants and uncover innovations in these enzymes associated with the evolution of endosperm, we queried genome and transcriptomes to identify NRPB, NRPD, and NRPE subunits from species representing the breadth of the land plant phylogeny (supplementary table S1, Supplementary Material online). Using phylogenetic analysis, we describe the evolution of NRPD and NRPE subunits, and changes in the RNA Polymerase (RDR), Dicer-like (DCL), and Argonaute (AGO) families across plant evolution, including all four clades of gymnosperms. We uncover the ancient origin of Pol V, as well as angiosperm-specific and seed plant-specific subunits of Pol IV and V, respectively. These findings indicate a more ancient origin of Pol V than previously described, and suggest that RdDM could be functional in the earliest land plants. Our data are also consistent with Escape from Adaptive Conflict (EAC) as the driving force behind the evolution of Pol IV and Pol V subunits.

Results

Identification of RdDM Machinery in Nonflowering Plants

To understand the duplication and specialization of proteins in the RdDM pathway, we identified putative orthologs for each unique Pol II/IV/V subunit (the first-, second-, fourth-, fifth-, and seventh largest subunits), as well as RDR, DCL, and AGO families from taxa across the land plant lineage (fig. 1). Using the eudicot *Arabidopsis thaliana* as a reference, we searched the fully sequenced genomes of the monocot *Oryza sativa*, the early-diverging angiosperm *Amborella trichopoda*, the lycophyte *Selaginella moellendorfii*, and the moss *Physcomitrella patens*. To expand the range of taxa and gain resolution on the tree of life, we also searched transcriptomes for the gymosperm *Gingko biloba*, the ferns

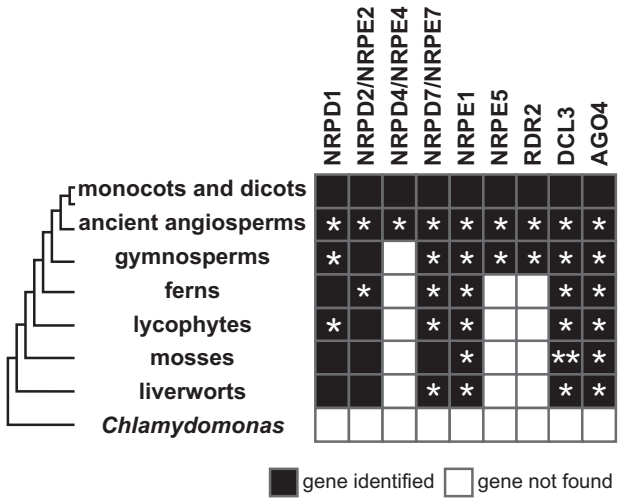

FIG. 1. Summary of species and genes analyzed in this study. (Left) Cladogram and list of land plant lineages searched in this study. (Right) Chart of gene presence. Filled boxes indicate identification of an ortholog of the gene listed at top. Single asterisk (*) represents genes first reported in this study; the double asterisk (**) has a described mutant (Cho et al. 2008), but no previously published sequence or phylogenetic characterization.

Pteridium aquilinum, *Ceratopteris richardii*, and *Pteris vittata*, and the liverwort *Marchantia paleacea* (see supplementary table S1, Supplementary Material online, for full list of databases and libraries).

We were particularly interested in gymnosperms, the closest relatives of angiosperms, which are poorly represented among available data sets. We therefore performed RNAseq on *Ephedra trifurca*, *Pinus canariensis*, and *Cycas revoluta*, generating 5–7.4 billion nucleotides of sequence, resulting in 40,000–50,000 open reading frames for each species (supplementary table S2, Supplementary Material online). Reverse transcription PCR and Sanger sequencing confirmed the sequence of gymnosperm orthologs, and where necessary cDNA ends were determined with RLM-RACE to resolve full-length coding sequences. During this analysis, the transcriptome of the gymnosperm *P. abies* was published (Nystedt et al. 2013); however, coverage of the Pol II, Pol IV, and Pol V subunits was incomplete in this species, and no inference on polymerase structure could be drawn.

RNA Pol V Is Present in All Land Plants

Because the largest subunit forms the catalytic center of the polymerase, in addition to encoding a distinctive CTD, the presence of NRPD1 and NRPE1 in a plant is indicative of Pol IV and Pol V activities, respectively. Although fragments of NRPD1 were recovered from a number of early-diverging plants, NRPE1 has been identified only in angiosperms (Luo and Hall 2007), suggesting that it might be responsible for the highly active RdDM found in flowering plants.

We identified clear orthologs of NRPD1 and NRPE1 in the first diverging lineage of angiosperms, *Am. trichopoda*, and all four gymnosperm species, indicating that Pol V was present in the earliest angiosperms and gymnosperms (fig. 2). Each of these orthologs contain the same domain structure as their *A. thaliana* counterparts, including WG/GW repeats and a DeCL domain in the CTD, further supporting their assignment as NRPE1 orthologs (supplementary fig. S1, Supplementary Material online).

We also identified NRPD1 and NRPE1 orthologs in fern, lycophyte, moss, and liverwort. Although the NRPD1 clade is well resolved, NRPE1 sequences do not form a monophyletic group. However, all of the NRPE1 sequences retrieved contain characteristic WG/GW-rich regions in the CTD indicating their orthology (fig. 2 and supplementary fig. S1, Supplementary Material online). Most NRPE1 orthologs also contain a DeCL domain as identified through sequence similarity with *A. thaliana* and through domain searches of the Pfam database. Surprisingly, the single NRPD1 homolog and both NRPE1 homologs in moss lack the DeCL domain (supplementary fig. S1, Supplementary Material online). However, the DeCL domain is found in *M. paleacea* NRPD1 and NRPE1 sequences, indicating that it was likely present in the ancestor of mosses and liverworts before being lost in *Ph. patens* and possibly other mosses. DNA methylation associated with repetitive sequences and larger (~23 nt) siRNAs have been reported in moss (Cho et al. 2008), suggesting a functional RdDM pathway. It is possible that the DeCL domain is dispensable in moss, or that it is present on a separate protein that associates with the Pol IV and Pol V holoenzyme complexes.

Composition of RNA Pol IV and Pol V Holoenzymes

In addition to a unique largest subunit, Pol IV and Pol V differ from Pol II in their second, fourth, fifth, and seventh subunits (NRPD2/E2, NRPD4/E4, NRPE5, and NRPD7/E7, respectively). Together with NRPD1 or NRPE1, NRPD2/E2 forms the catalytic region of the polymerases. This subunit has been reported in liverworts, suggesting that it is present in all land plants (Luo and Hall 2007). As expected, we identified NRPD2 in each of the species searched (supplementary fig. S2, Supplementary Material online).

In yeast, the fourth and seventh largest subunits work together as a subcomplex that is capable of dissociating from the catalytic subunits and regulating the transcribed mRNA (Ream et al. 2013). Interestingly, a mutation in the seventh subunit of *Schizosaccharomyces pombe* Pol II (Rpb7) has a defect in silencing without altering other Pol II activities (Djupedal et al. 2005), highlighting that this subcomplex might play an important role in the silencing function of Pol II and its duplication might be important for partitioning this activity. Previous work identified NRPD7/E7 but not NRPD4/E4 in moss (Tucker et al. 2010), indicating that these subunits duplicated at different times in the evolution of land plants. We identified orthologs of NRPD7/E7 in each of the species we queried, including *M. paleacea*, suggesting that NRPD7/E7 coevolved with NRPD1, NRPE1, and NRPD2/E2 (supplementary fig. S3, Supplementary Material online). However, an ortholog of NRPD4/E4 was absent from all of the fern and gymnosperm transcriptomes, but was identified in the earliest diverging angiosperm, *Am. trichopoda* (fig. 3). This result indicates the presence of a fourth subunit dedicated to silencing is an angiosperm innovation and might be responsible for the increased activity of Pol IV and Pol V in angiosperms.

Pol IV and Pol V share NRPD2/E2, NRPD4/E4, and NRPD7/E7, while Pol V additionally uses a specific fifth subunit, NRPE5. Recent evidence suggests that NRPE5 can also associate with Pol IV in *Zea mays* (Haag et al. 2014). We identified NRPE5 in all of the angiosperm and gymnosperm species tested, but failed to find an ortholog in any of the three fern transcriptomes, suggesting that this subunit arose after seed plants (angiosperms and gymnosperms) diverged from Pteridophytes (fig. 4).

Identifying Core Silencing Machinery in Land Plants

RNA Pol IV and Pol V do not act alone, but are part of the larger RdDM pathway, which includes specialized members of the RDR, DCL, and AGO families. Each of these families predates the divergence of plants and animals (Ahlquist 2002; Meister 2013; Wilson and Doudna 2013); however, the specific family members involved in RdDM are plant specific.

Arabidopsis thaliana contains ten Argonautes arranged into three clades (Vaucheret 2008), of which members of the

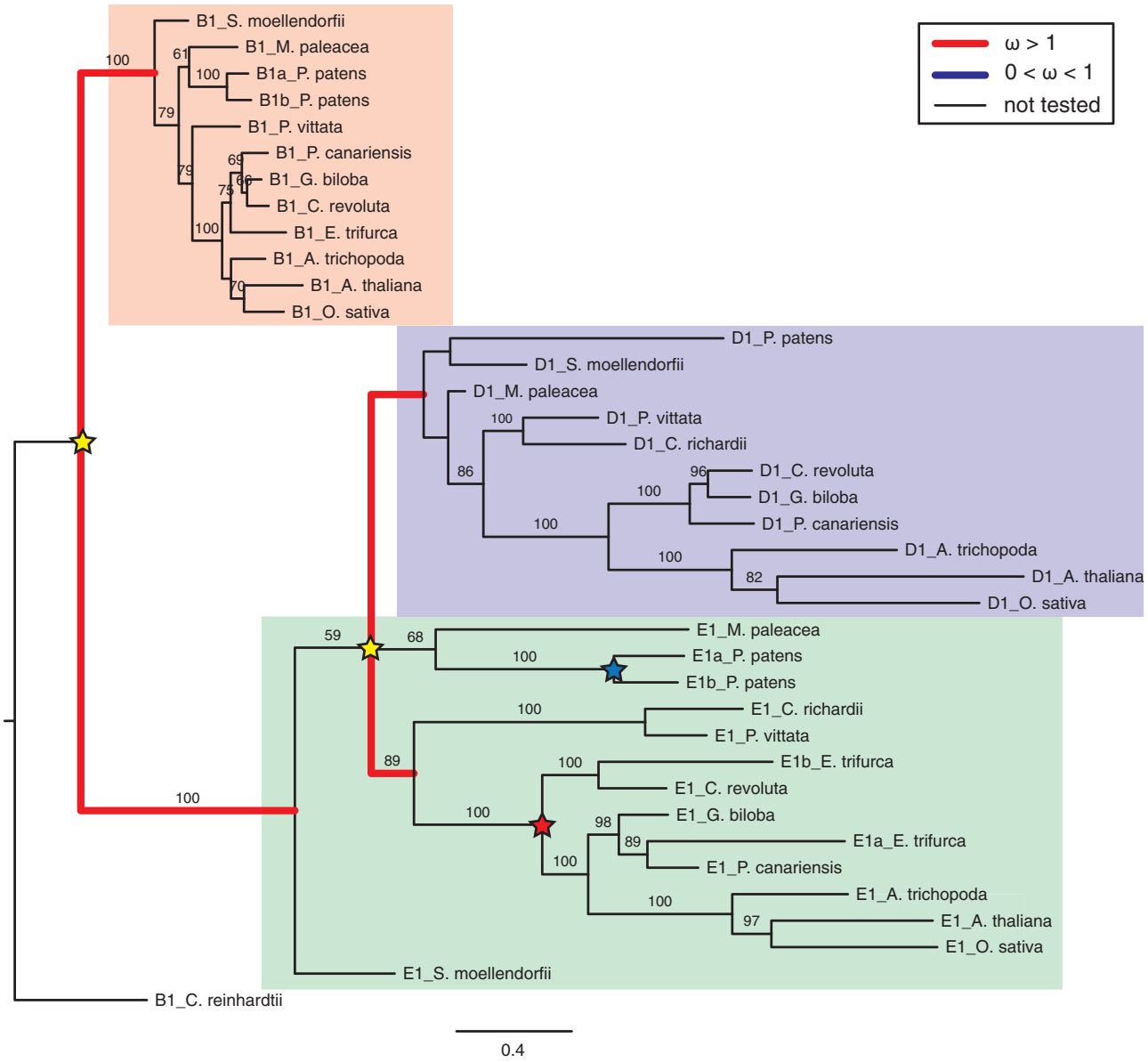

Fig. 2. NRPE1 is encoded in all land plants. A maximum-likelihood phylogenetic tree of NRPB1, NRPD1, and NRPE1 homologs demonstrates that orthologs of NRPD1 and NRPE1 exist in each group of land plants, including the liverwort *M. paleacea*. The *S. moellendorfii* NRPE1 falls in an unannotated region of the genome covered by two different scaffolds and therefore likely contains sequencing errors that cause its placement basal to the NRPD1/NRPE1 duplication. Yellow stars mark duplications that gave rise to distinct polymerases. Red and blue stars mark additional lineage-specific and species-specific duplications, respectively. Bootstrap support values ≥50 are listed on the branches. Thick red lines correspond to branches demonstrating positive selection ($\omega > 1$), as determined by the branch-sites test in PAML (Yang 2007). Only clades that were resolved in the mostly likely tree were tested.

AGO4 clade are associated with RdDM. More specifically, AGO4 plays the largest role in RdDM, but other clade members (AGO6, AGO8, and AGO9) might be partially redundant or have roles in specific tissues (Havecker et al. 2010; Olmedo-Monfil et al. 2010; Eun et al. 2011). Of the four DCLs in *A. thaliana*, only DCL3 generates the 24 nt siRNAs that trigger RdDM (Xie et al. 2004; Kasschau et al. 2007). An initial report indicated that DCL3 might be absent from conifers (Dolgosheina et al. 2008) and a DCL-like gene identified in larch was not phylogenetically clustered with other DCL3s, indicating that it might not be an ortholog (Zhang et al. 2013). DCL3-like genes are reported in lycophytes and moss (Cho et al. 2008; Axtell 2013), but phylogenetic analyses are lacking, and thus whether they are orthologs of *A. thaliana* DCL3 or represent less specialized members of the DCL family is unknown. Finally, three RDRs are involved in small RNA biogenesis in *A. thaliana*, but only RDR2 is required for the generation of p4-siRNAs (Xie et al. 2004; Kasschau et al. 2007). An RDR2-like gene is reported in lycophytes (Axtell 2013), but there is no phylogenetic analysis that demonstrates clear orthology.

To investigate the specialization of RNA silencing components for RdDM among land plants, we retrieved all sequences similar to RDR, DCL, and AGO families from transcriptome and genome databases and generated

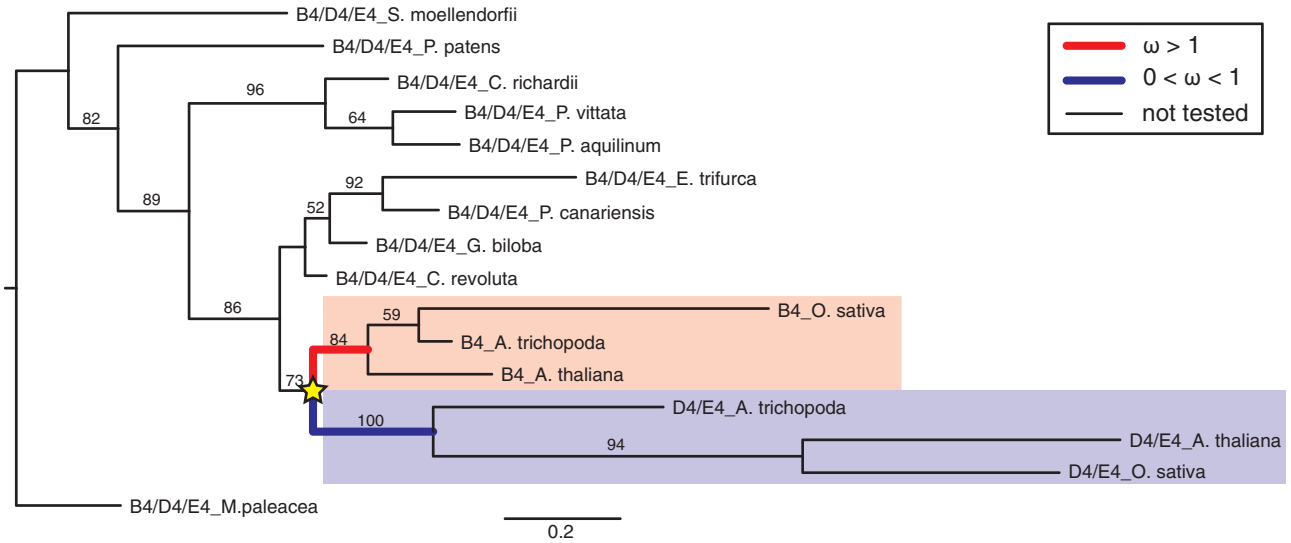

Fig. 3. NRPD4/E4 is angiosperm specific. A maximum-likelihood phylogenetic tree of NRPB4 and NRPD4/E4 homologs demonstrates that orthologs of NRPD4/E4 exist in the angiosperm *Am. trichopoda*, as well as angiosperms *A. thaliana* and *O. sativa*. NRPD4/E4 orthologs were not found in gymnosperms. The yellow star marks the duplication that gave rise to NRPD4/E4. The *Chlamydomonas reinhardtii* sequence was too divergent to be placed on the tree, so *M. paleacea* was used as a root. Bootstrap support values ≥50 are listed on the branches. Thick red lines correspond to branches demonstrating positive selection ($\omega > 1$), as determined by the branch-sites test in PAML (Yang 2007). Thick blue lines do not show a signature of positive selection immediately postduplication.

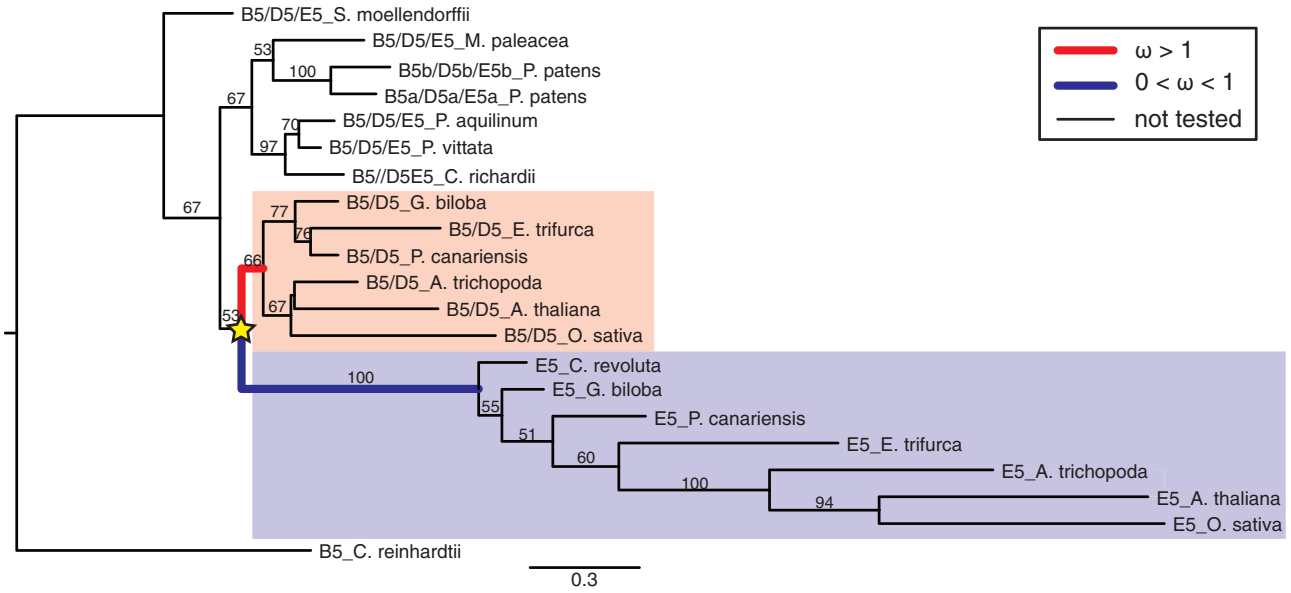

Fig. 4. NRPE5 is seed plant specific. A maximum-likelihood phylogenetic tree of NRPB5/D5 and NRPE5 homologs supports the duplication leading to NRPE5 (yellow star) between ferns and gymnsoperms. NRPE5 was identified in each of the four gymnosperm species tested and was not identified in any of the three fern species. Bootstrap support values ≥50 are listed on the branches. Thick red lines correspond to branches demonstrating positive selection ($\omega > 1$), as determined by the branch-sites test in PAML (Yang 2007). Thick blue lines do not show a signature of positive selection immediately postduplication.

maximum-likelihood trees to determine orthology and paralogy. Clear orthologs of DCL3 and AGO4 were identified in each species queried (supplementary figs. S4 and S5, Supplementary Material online), indicating that the duplications responsible for these specialized proteins occurred prior to the evolution of land plants. Because AGO4 binds the 24 nt siRNAs generated by DCL3, it is not surprising that these duplications appear to have coevolved.

Interestingly, land plant lineages from liverworts to ferns possess only two RDRs—an RDR6 ortholog and an RDR protein that is sister to both RDR1 and RDR2 (fig. 5). Only in seed plants was a distinct RDR2 ortholog detected, indicating that there might be no RDR specialized for p4-siRNA production in early diverging land plant lineages. Alternatively, the RDR1/2 homolog might function specifically in RdDM, and early diverging land plants might lack the antiviral activity of

RDR1. However, because antiviral small RNA defense is conserved in animals (Ding and Voinnet 2007), it is more parsimonious that RDR1 activity is ancestral and RDR2 function is derived.

Ongoing Duplication in RdDM Protein Families

In *A. thaliana*, NRPD7/E7 has duplicated and subfunctionalized into Pol IV- and Pol V-specific subunits (Ream et al. 2009; Tucker et al. 2010). A similar process likely occurs in *Z. mays*, which contains three NRPD2/E2 isoforms that are preferentially incorporated into unique polymerases, and a Pol IV/V-specific ninth subunit (Stonaker et al. 2009; Haag et al. 2014). *Arabidopsis thaliana* also encodes multiple copies of NRPB3, NRPB6, NRPB8, and NRPB9 (Ream et al. 2009). These do not appear to define polymerases with novel functions, but rather the duplicates may have been retained by subfunctionalization to generate polymerases with a subset of the ancestral functions (Tan et al. 2012).

In addition to the key duplications that gave rise to Pol IV- and Pol V-specific subunits, we also identified subsequent duplication events. Two putative NRPE1 subunits were identified in *E. trifurca*, suggesting that further duplication of the largest subunit is ongoing, a hypothesis supported by the phylogenetic placement of the *Cy. revoluta*

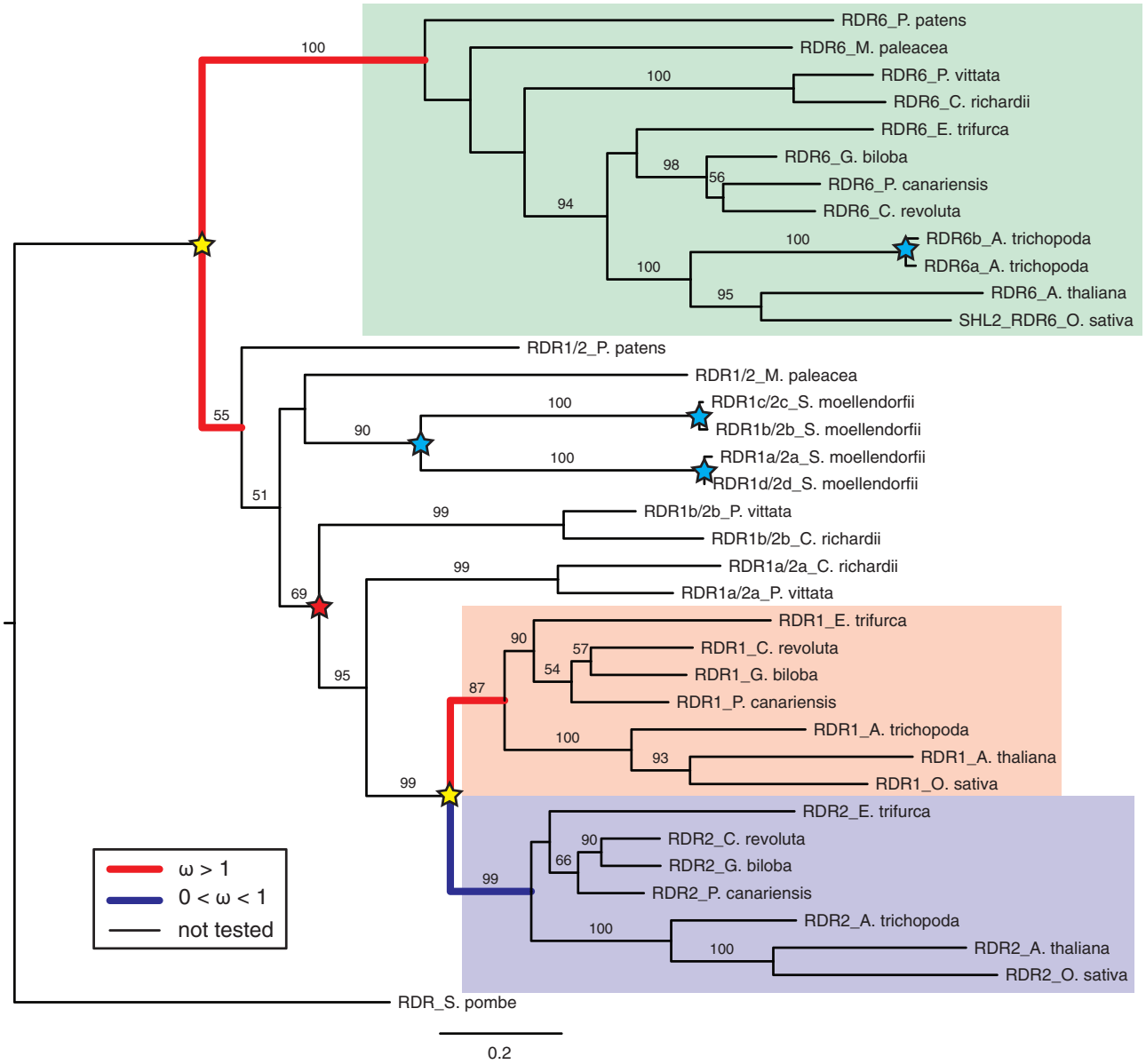

Fig. 5. RDR2 is seed plant specific. A maximum-likelihood phylogenetic tree of RDR homologs establishes that orthologs of RDR6 exist in all land plants, including the liverwort *Marchantia paleacea*; however, orthologs of RDR2 were identified only in seed plants (gymnosperms and angiosperms). Earlier diverging plant lineages encode an RDR ortholog that is sister to both RDR1 and RDR2. Yellow stars mark the duplication that gave rise to the three clades of RDR seen in seed plants. Red and blue stars mark additional lineage-specific and species-specific duplications, respectively. Because of a lack of RDR sequence in *Chlamydomonas reinhardtii*, *Schizosaccharomyces pombe* was used to root the tree. Bootstrap support values ≥50 are listed on the branches. Thick red lines correspond to branches demonstrating positive selection ($\omega > 1$), as determined by a likelihood ratio test in PAML (Yang 2007). Thick blue lines do not show a signature of positive selection immediately postduplication.

NRPE1 sister to the second *E. trifurca* sequence (fig. 2). Two NRPE1 sequences were also identified in *Ph. patens*. Additional species-specific (*Ph. patens*) and lineage-specific (fern) duplications were detected in NRPD2/E2 (supplementary fig. S2, Supplementary Material online). The AGO, DCL, and RDR families also appear to be subject to repeated duplication, because many species-specific and lineage-specific paralogs were recovered (fig. 5 and supplementary figs S4 and S5, Supplementary Material online). Each of the RdDM genes analyzed here has reverted to a single copy following the recent whole-genome triplication at the base of the Brassiceae (Huang et al. 2013), suggesting that duplication does not confer advantage simply through increased dosage of these proteins. Further phylogenetic and molecular analyses are needed to identify any sub- or neofunctionalization associated with these duplications.

Evolution of RNA Pol IV and Pol V Subunits

Small RNA-mediated transcriptional silencing occurs in many eukaryotes, including yeast, flies, worms, and mammals (Moazed 2009; Cecere and Grishok 2014). In fungi and metazoans, Pol II initiates small RNA biogenesis and produces scaffold transcripts that are bound by siRNA/AGO (or piRNA/PIWI) complexes. It is therefore likely that the ancestral eukaryotic Pol II performs the functions of plants' Pol II, Pol IV, and Pol V, and suggests that EAC explains the retention of duplicate polymerase subunits in plants. The EAC model proposes that a protein with multiple functions experiences adaptive conflict whereby constraint from one function limits the evolutionary optimization of the second function and vice versa (Hughes 1994; Hittinger and Carroll 2007; Marais and Rausher 2008). The two functions are locked in a tug-of-war before duplication allows subfunctionalization and each paralog is free to evolve unconstrained.

A key prediction of EAC is that both paralogs will undergo positive selection immediately postduplication as the adaptive conflict is resolved (Hughes 1994; Hittinger and Carroll 2007; Marais and Rausher 2008). This is in contrast to neofunctionalization, during which one paralog evolves a novel function by positive selection, while the other performs the ancestral function and remains under purifying selection. We performed the branch-sites test in PAML (Yang 2007) on each branch of a polymerase subunit family that had undergone duplication to determine whether there was evidence of positive selection postduplication (supplementary table S3, Supplementary Material online). As predicted by the EAC model, positive selection was detected on both branches following duplication of the largest subunit (fig. 2). However, for the second, fourth, fifth, and seventh subunits, positive selection was detected only on the branches subtending Pol II subunits (figs. 3 and 4 and supplementary figs. S2 and S3, Supplementary Material online).

A lack of positive selection on Pol IV- and Pol V-specific branches was also surprising given the long branch lengths associated with NRPD/E clades compared with the cognate NRPB clades (figs. 2–4 and supplementary figs. S2 and S3, Supplementary Material online). To investigate this systematically, we calculated patristic distances (sum of branch lengths) between all pairs of B, D, and E subunits and found branch lengths to be uniformly longer for all Pol IV and V subunits (fig. 6 and supplementary fig. S6, Supplementary Material online), suggesting on-going diversification in these proteins.

The core transcriptional silencing machinery (RDR, DCL, AGO) is also associated with post-transcriptional silencing and antiviral defense in diverse eukaryotes (Ding and Voinnet 2007; Jinek and Doudna 2009), suggesting that the ancestral proteins had multiple functions. To determine if the EAC model might explain duplication and specialization

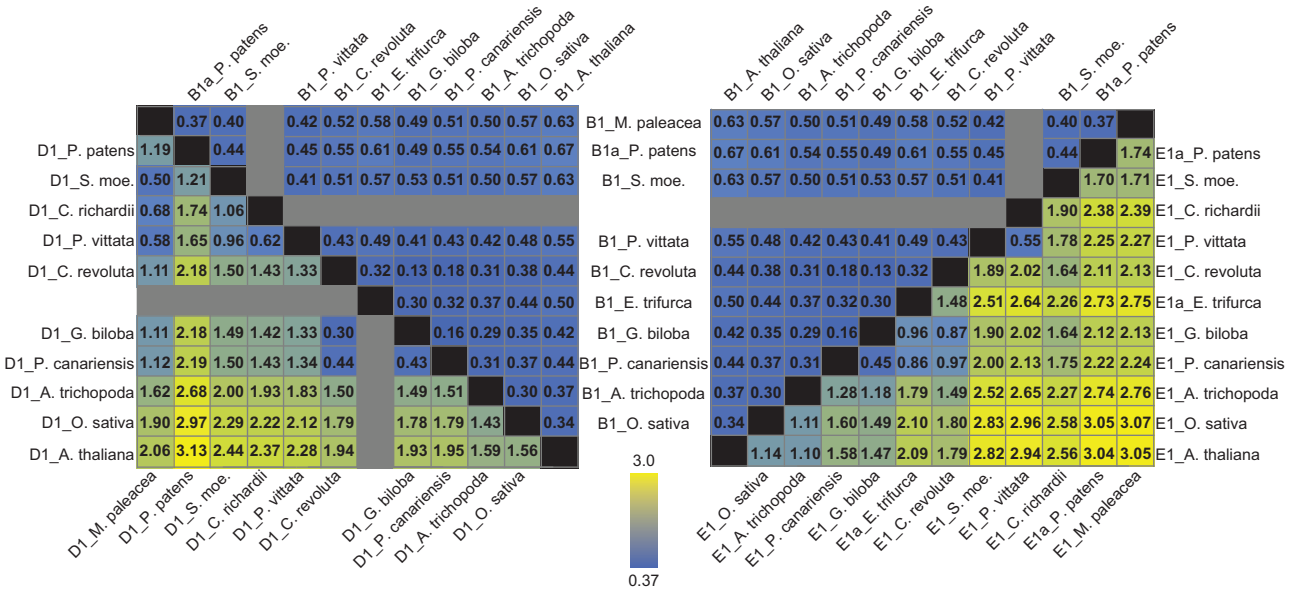

Fig. 6. NRPD1 and NRPE1 have higher patristic distances than NRPB1. A heat map of patristic distances (sum of branch lengths) for all combinations of NRPB1, NRPD1, and NRPE1 show greater divergence for NRPD1 (bottom left) and NRPE1 (bottom right), compared with NRPB1 (top middle). Branch lengths were calculated from the most likely tree.

among these gene families in plants, we again used the branch-sites test of PAML (Yang 2007) on the RDR, DCL, and AGO duplications (supplementary table S3, Supplementary Material online). As expected, there is evidence for positive selection on both branches following the duplications that give rise to DCL3 and AGO4 (supplementary figs. S4 and S5, Supplementary Material online). Similarly, both branches display positive selection following the duplication that gives rise to RDR6 and RDR1/2 (fig. 5). However, the subsequent duplication that separates the RDR1 and RDR2 functions shows positive selection only on the branch subtending RDR1. In *A. thaliana* and maize, RDR2 physically associates with Pol IV, and might therefore be considered part of the larger Pol IV complex (Haag et al. 2012, 2014). A lack of positive selection on the RDR2 branch therefore follows the pattern of smaller Pol IV subunits, which show positive selection only on the opposite branch.

Rapid Divergence in the NRPE1 CTD

Although the relatively long branch lengths within catalytic regions of NRPD1 and NRPE1 are striking, more noteworthy is the rapid divergence within the NRPE1 CTDs, which have no identifiable sequence homology outside the DeCL domain (supplementary fig. S1, Supplementary Material online). Between the catalytic region and the DeCL domain, NRPE1 orthologs contain a WG/GW platform, a repeat region rich in WG, GW, and GWG peptides also known as AGO hooks (El-Shami et al. 2007; Till et al. 2007). This region is responsible for mediating the interaction between the Pol V CTD and AGO4 (El-Shami et al. 2007) and similar WG/GW platforms mediate AGO association in diverse proteins across eukaryotes (Till et al. 2007; Bednenko et al. 2009; Bies-Etheve et al. 2009; Karlowski et al. 2010). The WG/GW platform is known to be divergent among angiosperm NRPE1 (El-Shami et al. 2007), but repeats in this region have not been thoroughly analyzed.

We used RADAR and BLAST to identify repeat units in each of the NRPE1 CTDs (fig. 7 and supplementary fig. S7, Supplementary Material online) and identified several interesting aspects of these repeats. First, although nearly all CTDs possess degenerate direct repeats that include AGO hook peptides, there are a few notable exceptions. *Physcomitrella patens* NRPE1a and NRPE1b possess AGO hook peptides, but lack any detectable repeats (fig. 7) and the five repeats in *E. trifurca* NRPE1b lack any AGO hook peptides (supplementary fig. S7, Supplementary Material online). This implies that it is the presence of AGO hook peptides rather than the direct repeats that is of functional significance to the protein.

It is also notable that AGO hook peptides do not occur at similar frequencies in all WG/GW platforms. The amino acids in these peptides range from $<2.5\%$ of the region between domain H and the DeCL (*Ph. patens* NRPE1a) to $>10\%$ of this region (*Am. trichopoda* NRPE1). Some taxa also display a preference for one type of AGO hook peptide within the WG/GW platform. *Arabidopsis thaliana* NRPE1 contains 17 WGs, but only 1 GW and 1 GWG, while the *Cy. revoluta* NRPE1 contains 20 GWs and no WG or GWGs. *Cycas revoluta* NRPE1 even lacks the conserved WG found at the end of domain H (fig. 7). Because many of these peptides are not part of the repeat unit, this difference is unlikely to be a consequence of different repeat sequences, but might rather suggest a difference in preference for AGO4/NPRE1 association.

Alignment of repeat copies in each protein reveals a dramatic difference in length and sequence of repeats, as well as in the level of conservation between repeat units (fig. 7 and supplementary fig. S7, Supplementary Material online). This indicates that there is frequent and repeated expansion of sequences within the platform to create new repeats. Indeed several NRPE1 orthologs contain multiple distinct repeat units supporting the hypothesis of recurrent rounds of expansion (fig. 7 and supplementary fig. S7, Supplementary Material online). Steady deterioration of these repeats due to lack of selection might account for the presence of AGO hook peptides outside of detectable repeats. Three of the NRPE1 orthologs also contain regions of simple repeats: *A. thaliana* NRPE1 has a QS-rich region, while *Pi. canariensis* NRPE1 and *E. trifurca* NRPE1a have GR-rich repeats. No function has been ascribed to these simple repeats, suggesting that they might be bystanders of the illegitimate recombination that likely drives duplication within the WG/GW platform (Kane et al. 2010).

Discussion

Our analysis demonstrates that RNA Pol V is an ancient polymerase present in all land plants and not restricted to angiosperms as has been suggested previously (Luo and Hall 2007; Tucker et al. 2010; Matzke and Mosher 2014). Furthermore, the core RdDM machinery necessary to produce and utilize p4-siRNAs is also present in the earliest diverging plant lineages. It was reported that the loss of NRPD2 might account for the large, unmethylated genomes commonly found in gymnosperms (Lee et al. 2011). However, our analysis suggests that NRPD2 is present in all gymnosperms and recent publications indicate that p4-siRNAs are produced in conifers (Wan et al. 2012; Nystedt et al. 2013; Zhang et al. 2013), hinting at a functional gymnosperm Pol IV.

Although Pol IV and V subunits are present in the earliest land plant lineages, analysis of smaller subunits within these holoenzymes and careful phylogenetic assessment of the RDR family suggests that innovations in RdDM proteins occurred later and are associated with the evolution of seeds and flowers. The duplications creating NRPD and NRPE first, second, and seventh subunits occurred prior to the evolution of extant land plant lineages, suggesting that all plants have distinct Pol II, Pol IV, and Pol V holoenzymes. In nonflowering seed plants (gymnosperms), Pol V gains a specific fifth subunit and the function of Pol IV is potentially altered by the evolution of a dedicated RDR2. Finally, in flowering plants, both Pol IV and V develop a specialized fourth subunit. Further research is needed to confirm the assembly of these ancient polymerases and to assess the functional consequences of these recent duplications. It will be particularly interesting to determine if biochemical changes in Pol IV or Pol V account for the high level of p4-siRNAs observed in angiosperms.

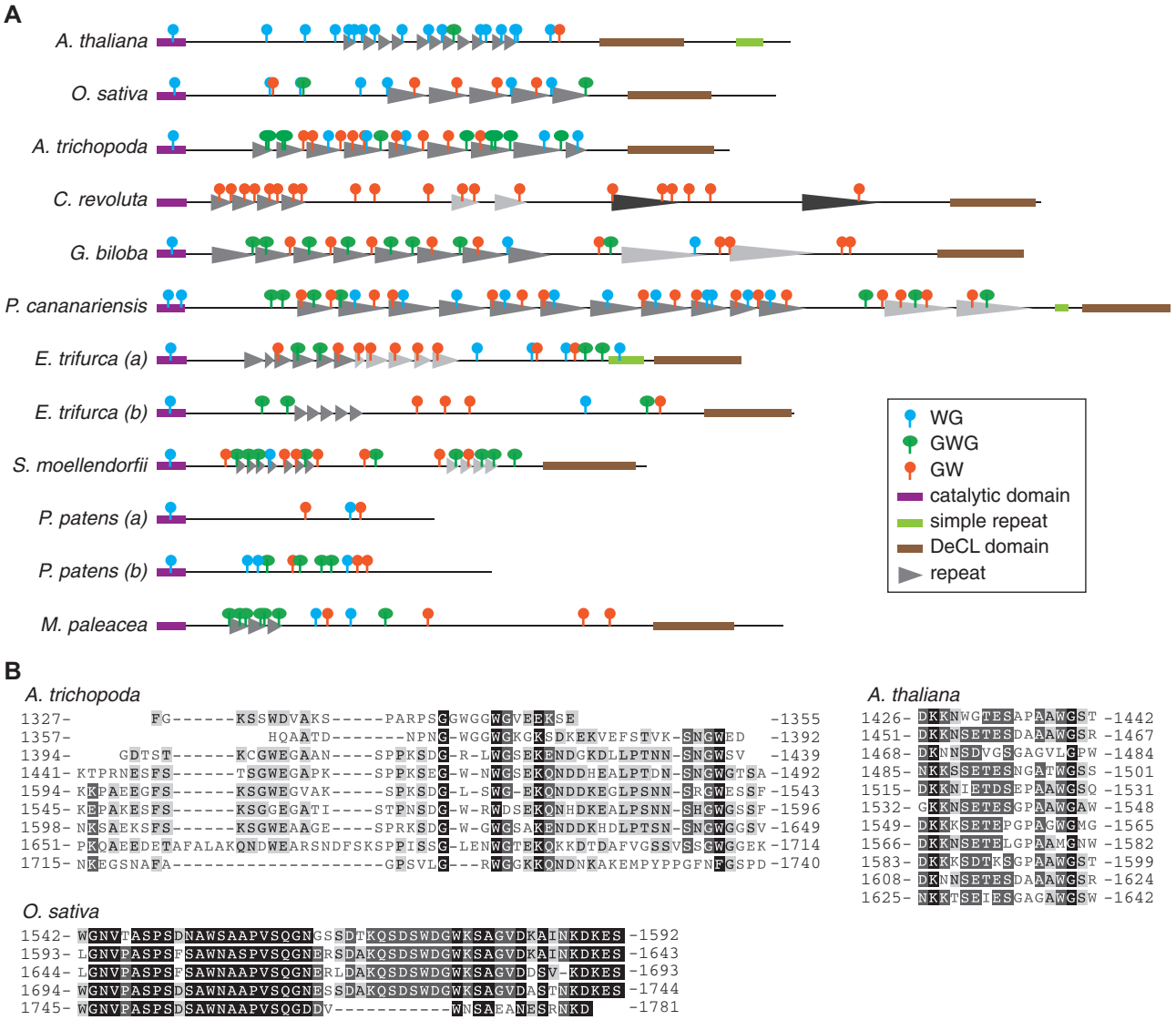

Fig. 7. Variability in NRPE1 CTDs. (*A*) Diagrams of the C-terminal domains of NRPE1 orthologs showing variation in length, repeat number and sequence, and number and position of GW/WG/GWG peptides. (*B*) Alignments of repeat units in three angiosperm sequences showcase the diversity of repeat sequence, length, and conservation.

It is striking that six of the nine genes assessed here are present in the earliest diverging land plant lineages, opening the possibility that they may have duplicated simultaneously and that their respective retentions are interdependent. A partial NRPD1 sequence was identified within four species of algae in the family Characeae, a close ancestor of land plants (Luo and Hall 2007). NRPE1 and NRPD2/E2 were not identified in these species; however, the degenerate PCR used to search for them also failed to identify NRPE1 in all nonflowering plants tested (Luo and Hall 2007). To determine whether other Pol IV and Pol V subunits are present in Characeae or the wider grouping of *Streptophyta*, we searched the six largest assembled transcriptomes present in the 1KP project from this group of algae. We were unable to identify NRPE1-specific reads, but were also unable to assemble NRPB1, suggesting that the read depth of these transcriptomes is not sufficient for our purpose. We also searched the genome of the charophyte *Klebsormidium flaccidum* (Hori et al. 2014), but were unable to detect orthologs of NRPD1, NRPE1, NRPD2/E2, or NRPD7/E7. Further work is needed to determine to what extent the RdDM pathway exists in specific algal lineages and to assess the timing of the earliest NRPB to NRPD/E duplications. Beyond increasing our understanding of RdDM, determining the extent to which components of the different polymerases duplicated simultaneously versus sequentially will broaden our knowledge about the evolution of multi-subunit complexes.

EAC is the best model to account for subfunctionalization of gene transcription and silencing activities of ancestral Pol II subunits because there is evidence that Pol II functions during transcriptional silencing in diverse eukaryotes (Moazed 2009; Cecere and Grishok 2014). However, it is possible that transcriptional silencing evolved independently in plants through neofunctionalization of NRPD/E subunits following duplication. Although our data are agnostic as to whether

transcriptional silencing evolved once or independently in multiple lineages, they do not support a model of neofunctionalization after duplication because there is positive selection on all NRPB branches following duplication. The lack of positive selection on most NRPD/E branches might suggest that selection for Pol II function is stronger than Pol IV or Pol V function or that our phylogenetic trees are not sufficiently dense surrounding the duplication events to detect selection on NRPD/E branches. Alternatively, the EAC model might not accurately describe evolution of multi-subunit complexes (Beilstein et al. 2015) and additional theoretical work (Sikosek et al. 2012) will be needed to model these complex interactions. Indeed, it is possible that incorporation into multiple complexes with distinct functions drives the adaptive conflict of smaller subunits.

One of the most striking findings in this study is the divergence among CTDs of NRPE1. Each NRPE1 ortholog contains a WG/GW platform in the CTD and nearly all these platforms contain embedded tandem direct repeats. However, there is no sequence identity between the WG/GW platforms, suggesting that these repeats undergo rounds of sequence deterioration and expansion through illegitimate recombination (Kane et al. 2010; Schaper et al. 2014). Indeed, the presence of highly similar repeats within a platform, such as *O. sativa* NRPE1, suggests that WG/GW platform expansion is ongoing. The rapid divergence in repeat sequence is in contrast to the large majority of human tandem repeat proteins, which show deep conservation of repeat sequence and structure (Schaper et al. 2014). It is unclear why tandem repeats would persist in the NRPE1 CTD over such evolutionary distance, particularly in the absence of sequence homology, but this observation suggests important functions for both the repeat structure and the rapid divergence of repeat sequence.

Materials and Methods

Ortholog Identification

Arabidopsis thaliana protein sequences were used as queries for TBLASTN or BLASTP searches against the *O. sativa*, *Am. trichopoda*, *S. moellendorffii*, and *Ph. patens* whole genomes (supplementary table S1, Supplementary Material online). Where putative orthologs fell into unannotated regions, genes were predicted from genomic sequence using FGENESH (www.softberry.com, last accessed March 18, 2015). RNA-seq reads from *G. biloba* were downloaded and assembled with Trinity (Grabherr et al. 2011). Assembled transcriptomes of *G. biloba*, *Pt. aquilinum*, *Pte. vittata*, *C. richardii*, and *M. paleacea* were searched with TBLASTN using *A. thaliana* protein sequences as queries. In some cases, multiple nonoverlapping open reading frames were manually linked. Partial sequences of meaningful length were retained for further analysis. Geneious 6.1.8 identified protein domains with InterProScan of the Pfam database. Predicted orthologs that lacked conserved domains (but not partial transcripts) were excluded from further analysis.

Tissue Collection, RNA Extraction, and Sequencing of Gymnosperms

Tissues of *Cy. revoluta* and *Pi. canariensis* were collected from the University of Arizona Campus Arboretum. *Ephedra trifurca* samples were collected from a natural habitat in Tucson, AZ. Total RNA was extracted by homogenizing tissues in CTAB buffer (2% Cetyltrimethylammonium Bromide [CTAB], 100 mM Tris [pH 8], 2 M NaCl, 25 mM ethylenediaminetetraacetic acid [EDTA], 2% w/v polyvinylpyrrolidone [PVP], 0.2% β-mercaptoethanol). After a 10-min 65°C incubation, samples were extracted twice with 25:24:1 phenol:chloroform:isoamyl alcohol and once with 24:1 chloroform:isoamyl alcohol before precipitation with sodium acetate and ethanol.

Total RNA was submitted to the University of Missouri Columbia DNA Core Facility for Illumina RNA-seq library preparation and sequencing. RNA-seq reads were preprocessed by cutadapt 1.0 (Martin 2011) and PRINSEQ (Schmieder and Edwards 2011) before de novo transcriptome assembly with Trinity (Grabherr et al. 2011). Pol IV/V and other related genes were identified through TBLASTN of the resulting libraries.

For cDNA verification and RACE, RNA was extracted with the Spectrum Plant Total RNA kit (Sigma-Aldrich) and treated with DNA-free Turbo (Ambion) before reverse transcription with SuperScript III Reverse Transcriptase (Invitrogen) and amplification with Phusion DNA polymerase (New England Biolabs). Resulting products were cloned into pGEM-T (Promega) and sequenced.

Phylogenetic Analysis

Nucleotide sequences were aligned by translated amino acids using the MUSCLE algorithm (Edgar 2004) in Geneious version 6.1.8. (Biomatters Ltd., Auckland, New Zealand). Geneious was used to align, visualize, and manually correct alignments. Phylogenetic analysis was performed on full-length CDS alignments for most genes, and on conserved sequences for the largest subunit (catalytic regions B–H), RDR (RdRP domain), and AGO (PAZ-PIWI). Maximum-likelihood trees were inferred with RAxML version 7.2.8 (Stamatakis 2014) using a general time reversible model with gamma distributed rate heterogeneity. Support values for nodes in the tree were calculated from 100 bootstrap replicates. Trees were visualized with Figtree v1.4.0. (http://tree.bio.ed.ac.uk/software/figtree/, last accessed March 18, 2015) and poorly supported nodes were collapsed manually. Patristic distances were calculated in Geneious based on RAxML trees. Positive selection by the branch-sites test was inferred with PAML version 4.6 codeml (Yang 2007) on the iPlant Discovery Environment (Goff et al. 2011). Branches showing a significant signature of positive selection were detected by likelihood ratio test using X^2. The effect of different starting ω values on the calculation of total likelihood for each gene under the M1 model was explored for $\omega = 0.2$, 0.4, 0.6, 0.8, and 1.0. The stability of likelihood scores was determined by replicating each analysis a minimum of three times under the same model parameters. NRPE1 repeats were predicted with

RADAR (Neuwald 2009) or BLAST (bl2seq) (Altschul et al. 1990) and manually curated.

Supporting Data

The following sequences have been deposited at DNA Data Bank of Japan/EMBL/GenBank: Full-length confirmed cDNAs from *Cy. revoluta*, *G. biloba*, *E. trifurca*, and *Pi. canariensis*: KJ473663-KJ473694; RNAseq reads from *Cy. revoluta*, *E. trifurca*, and *Pi. canariensis*: SRR1525778, SRR1531150, SRR1531151; and Transcriptome Shotgun Assemblies from *Cy. revoluta*, *E. trifurca*, and *Pi. canariensis*: GBJU00000000, GBKT00000000, and GBLJ00000000.

Additional data are available from TreeBASE (study 16473 http://purl.org/phylo/treebase/phylows/study/TB2:S16473, last accessed March 18, 2015), including all nucleotide sequences, alignments, and tree files used in this study.

Supplementary Material

Supplementary figures S1–S7 and tables S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Steven E. Smith for assistance locating *E. trifurca* and Tanya Quist for access to the University of Arizona Campus Arboretum. Support for the generation of the *M. paleacea* transcriptome comes from UC MEXUS Collaborative program (grant 2011-UCMEXUS-19941-44-OAC7), Consejo Nacional de Ciencia y Tecnología (CONACYT) (grants CB-158550 and CB-158561), COSEAMX1 JEAI EPIMAIZE grant from the Institut de Recherche pour le Développement, and Universidad Veracruzana (Cuerpo Académico CA-UVER-234). This work was also supported by National Science Foundation grant MCB-1243608. *Pteris vittata* and *C. richardii* transcriptomes were generated by Jody Banks and Nadia Atallah and kindly shared prior to publication. We are also grateful to Michael Melkonian and Gane Wong for access to *Streptophyta* transcriptomes. Finally, we are indebted to the numerous plant genome projects funded by the National Science Foundation (the Amborella Genome Project, the Rice Genome Annotation Project, the Arabidopsis Information Resource), the Department of Energy (JGI and Phytozome), or the National Institutes of Health (the Medicinal Plant Genomics Resource).

References

Ahlquist P. 2002. RNA-dependent RNA polymerases, viruses, and RNA silencing. *Science* 296:1270–1273.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215:403–410.

Axtell MJ. 2013. Classification and comparison of small RNAs from plants. *Annu Rev Plant Biol.* 64:137–159.

Bednenko J, Noto T, DeSouza LV, Siu KW, Pearlman RE, Mochizuki K, Gorovsky MA. 2009. Two GW repeat proteins interact with *Tetrahymena thermophila* argonaute and promote genome rearrangement. *Mol Cell Biol.* 29:5020–5030.

Beilstein MA, Renfrew KB, Song X, Shakirov EV, Zanis MJ, Shippen DE. 2015. Evolution of the telomere-associated protein POT1a in *Arabidopsis thaliana* is characterized by positive selection to reinforce protein-protein interaction. *Mol Biol Evol.* 32:1329–1341.

Bies-Etheve N, Pontier D, Lahmy S, Picart C, Vega D, Cooke R, Lagrange T. 2009. RNA-directed DNA methylation requires an AGO4-interacting member of the SPT5 elongation factor family. *EMBO Rep.* 10:649–654.

Cecere G, Grishok A. 2014. A nuclear perspective on RNAi pathways in metazoans. *Biochim Biophys Acta.* 1839:223–233.

Cho SH, Addo-Quaye C, Coruh C, Arif MA, Ma Z, Frank W, Axtell MJ. 2008. Physcomitrella patens DCL3 is required for 22-24 nt siRNA accumulation, suppression of retrotransposon-derived transcripts, and normal development. *PLoS Genet.* 4:e1000314.

Ding SW, Voinnet O. 2007. Antiviral immunity directed by small RNAs. *Cell* 130:413–426.

Djupedal I, Portoso M, Spahr H, Bonilla C, Gustafsson CM, Allshire RC, Ekwall K. 2005. RNA Pol II subunit Rpb7 promotes centromeric transcription and RNAi-directed chromatin silencing. *Genes Dev.* 19:2301–2306.

Dolgosheina E, Morin RD, Aksay G, Sahinalp S, Magrini V, Margossian ER, Mattsson J, Unrau PJ. 2008. Conifers have a unique small RNA silencing signature. *RNA* 14:1508–1515.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

El-Shami M, Pontier D, Lahmy S, Braun L, Picart C, Vega D, Hakimi MA, Jacobsen SE, Cooke R, Lagrange T. 2007. Reiterated WG/GW motifs form functionally and evolutionarily conserved ARGONAUTE-binding platforms in RNAi-related components. *Genes Dev.* 21: 2539–2544.

Eun C, Lorkovic ZJ, Naumann U, Long Q, Havecker ER, Simon SA, Meyers BC, Matzke MA. 2011. AGO6 functions in RNA-mediated transcriptional gene silencing in shoot and root meristems in *Arabidopsis thaliana*. *PLoS One* 6:e25730.

Goff SA, Vaughn M, McKay S, Lyons E, Stapleton AE, Gessler D, Matasci N, Wang L, Hanlon M, Lenards A, et al. 2011. The iPlant Collaborative: Cyberinfrastructure for Plant Biology. *Front Plant Sci.* 2:34.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 29:644–652.

Haag JR, Brower-Toland B, Krieger EK, Sidorenko L, Nicora CD, Norbeck AD, Irsigler A, LaRue H, Brzeski J, McGinnis K, et al. 2014. Functional diversification of maize RNA polymerase IV and V subtypes via alternative catalytic subunits. *Cell Rep.* 9:378–390.

Haag JR, Ream TS, Marasco M, Nicora CD, Norbeck AD, Pasa-Tolic L, Pikaard CS. 2012. In vitro transcription activities of Pol IV, Pol V, and RDR2 reveal coupling of Pol IV and RDR2 for dsRNA synthesis in plant RNA silencing. *Mol Cell.* 48:811–818.

Havecker ER, Wallbridge LM, Hardcastle TJ, Bush MS, Kelly KA, Dunn RM, Schwach F, Doonan JH, Baulcombe DC. 2010. The *Arabidopsis* RNA-directed DNA methylation argonautes functionally diverge based on their expression and interaction with target loci. *Plant Cell* 22:321–334.

Henderson IR, Johnson L, Zhang X, Lu C, Meyers BC, Green PJ, Jacobsen SE. 2006. Dissecting *Arabidopsis thaliana* DICER function in small RNA processing, gene silencing and DNA methylation patterning. *Nat Genet.* 38:721–725.

Herr AJ, Jensen M, Dalmay T, Baulcombe DC. 2005. RNA polymerase IV directs silencing of endogenous DNA. *Science* 308:118–120.

Hittinger CT, Carroll SB. 2007. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* 449:677–681.

Hori K, Maruyama F, Fujisawa T, Togashi T, Yamamoto N, Seo M, Sato S, Yamada T, Mori H, Tajima N, et al. 2014. *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. *Nat Commun.* 5:3978.

Huang L, Jones AME, Searle I, Patel K, Vogler H, Hubner NC, Baulcombe DC. 2009. An atypical RNA polymerase involved in RNA silencing shares small subunits with RNA polymerase II. *Nat Struct Mol Biol.* 16:91–93.

Huang Y, Kendall T, Mosher R. 2013. Pol IV-dependent siRNA production is reduced in *Brassica rapa*. *Biology* 2:1210–1223.
Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci.* 256:119–124.
Jinek M, Doudna JA. 2009. A three-dimensional view of the molecular machinery of RNA interference. *Nature* 457:405–412.
Kane J, Freeling M, Lyons E. 2010. The evolution of a high copy gene array in *Arabidopsis*. *J Mol Evol.* 70:531–544.
Kanno T, Aufsatz W, Jaligot E, Mette MF, Matzke MA. 2005. A SNF2-like protein facilitates dynamic control of DNA methylation. *EMBO Rep.* 6:649–655.
Karlowski WM, Zielezinski A, Carrère J, Pontier D, Lagrange T, Cooke R. 2010. Genome-wide computational identification of WG/GW argonaute-binding proteins in *Arabidopsis*. *Nucleic Acids Res.* 38: 4231–4245.
Kasschau KD, Fahlgren N, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Carrington JC. 2007. Genome-wide profiling and analysis of *Arabidopsis* siRNAs. *PLoS Biol.* 5:e57.
Lahmy S, Pontier D, Cavel E, Vega D, El-Shami M, Kanno T, Lagrange T. 2009. PolV(PolIVb) function in RNA-directed DNA methylation requires the conserved active site and an additional plant-specific subunit. *Proc Natl Acad Sci U S A.* 106:941–946.
Lee EK, Cibrian-Jaramillo A, Kolokotronis SO, Katari MS, Stamatakis A, Ott M, Chiu JC, Little DP, Stevenson DW, McCombie WR, et al. 2011. A functional phylogenomic view of the seed plants. *PLoS Genet.* 7: e1002411.
Li CF, Pontes O, El-Shami M, Henderson IR, Bernatavichute YV, Chan SW, Lagrange T, Pikaard CS, Jacobsen SE. 2006. An ARGONAUTE4-containing nuclear processing center colocalized with Cajal bodies in *Arabidopsis thaliana*. *Cell* 126:93–106.
Luo J, Hall BD. 2007. A multistep process gave rise to RNA polymerase IV of land plants. *J Mol Evol.* 64:101–112.
Marais Des DL, Rausher MD. 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* 454:762–765.
Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17:10–12.
Matzke MA, Mosher RA. 2014. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat Rev Genet.* 15: 394–408.
Meister G. 2013. Argonaute proteins: functional insights and emerging roles. *Nat Rev Genet.* 14:447–459.
Moazed D. 2009. Small RNAs in transcriptional gene silencing and genome defence. *Nature* 457:413–420.
Morin RD, Aksay G, Dolgosheina E, Ebhardt HA, Magrini V, Mardis ER, Sahinalp SC, Unrau PJ. 2008. Comparative analysis of the small RNA transcriptomes of *Pinus contorta* and *Oryza sativa*. *Genome Res.* 18: 571–584.
Mosher RA. 2010. Maternal control of Pol IV-dependent siRNAs in *Arabidopsis* endosperm. *New Phytol.* 186:358–364.
Mosher RA, Melnyk CW, Kelly KA, Dunn RM, Studholme DJ, Baulcombe DC. 2009. Uniparental expression of PolIV-dependent siRNAs in developing endosperm of *Arabidopsis*. *Nature* 460:283–286.
Mosher RA, Schwach F, Studholme D, Baulcombe DC. 2008. PolIVb influences RNA-directed DNA methylation independently of its role in siRNA biogenesis. *Proc Natl Acad Sci U S A.* 105:3145–3150.
Neuwald AF. 2009. Rapid detection, classification and accurate alignment of up to a million or more related protein sequences. *Bioinformatics* 25:1869–1875.
Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, et al. 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature* 497:579–584.
Olmedo-Monfil V, Durán-Figueroa N, Arteaga-Vázquez M, Demesa-Arévalo E, Autran D, Grimanelli D, Slotkin RK, Martienssen RA, Vielle-Calzada JP. 2010. Control of female gamete formation by a small RNA pathway in *Arabidopsis*. *Nature* 464:628–632.
Onodera Y, Haag JR, Ream TS, Nunes P, Pontes O, Pikaard CS. 2005. Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* 120:613–622.
Pontier D, Yahubyan G, Vega D, Bulski A, Saez-Vasquez J, Hakimi MA, Lerbs-Mache S, Colot V, Lagrange T. 2005. Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in *Arabidopsis*. *Genes Dev.* 19:2030–2040.
Ream TS, Haag JR, Pikaard CS. 2013. Plant multisubunit RNA polymerases IV and V Nucleic acid polymerases. Vol. 30. Nucleic acids and molecular biology. (Berlin) Heidelberg: Springer. p. 289–308.
Ream TS, Haag JR, Wierzbicki AT, Nicora CD, Norbeck AD, Zhu JK, Hagen G, Guilfoyle TJ, Pasa-Tolić L, Pikaard CS. 2009. Subunit compositions of the RNA-silencing enzymes Pol IV and Pol V reveal their origins as specialized forms of RNA polymerase II. *Mol Cell.* 33: 192–203.
Schaper E, Gascuel O, Anisimova M. 2014. Deep conservation of human protein tandem repeats within the eukaryotes. *Mol Biol Evol.* 31: 1132–1148.
Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27:863–864.
Sidorenko L, Dorweiler JE, Cigan AM, Arteaga-Vazquez M, Vyas M, Kermicle J, Jurcin D, Brzeski J, Cai Y, Chandler VL. 2009. A dominant mutation in mediator of paramutation2, one of three second-largest subunits of a plant-specific RNA polymerase, disrupts multiple siRNA silencing processes. *PLoS Genet.* 5:e1000725.
Sikosek T, Chan HS, Bornberg-Bauer E. 2012. Escape from adaptive conflict follows from weak functional trade-offs and mutational robustness. *Proc Natl Acad Sci U S A.* 109:14888–14893.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.
Stonaker JL, Lim JP, Erhard KF, Hollick JB. 2009. Diversity of Pol IV function is defined by mutations at the maize rmr7 locus. *PLoS Genet.* 5: e1000706.
Tan EH, Blevins T, Ream TS, Pikaard CS. 2012. Functional consequences of subunit diversity in RNA polymerases II and V. *Cell Rep.* 1: 208–214.
Till S, Lejeune E, Thermann R, Bortfeld M, Hothorn M, Enderle D, Heinrich C, Hentze MW, Ladurner AG. 2007. A conserved motif in Argonaute-interacting proteins mediates functional interactions through the Argonaute PIWI domain. *Nat Struct Mol Biol.* 14: 897–903.
Tucker SL, Reece J, Ream TS, Pikaard CS. 2010. Evolutionary history of plant multisubunit RNA polymerases IV and V: subunit origins via genome-wide and segmental gene duplications, retrotransposition, and lineage-specific subfunctionalization. *Cold Spring Harb Symp Quant Biol.* 75:285–297.
Vaucheret H. 2008. Plant ARGONAUTES. *Trends Plant Sci.* 13:350–358.
Wan LC, Wang F, Guo X, Lu S, Qiu Z, Zhao Y, Zhang H, Lin J. 2012. Identification and characterization of small non-coding RNAs from Chinese fir by high throughput sequencing. *BMC Plant Biol.* 12:146.
Wierzbicki AT, Haag JR, Pikaard CS. 2008. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* 135:635–648.
Wierzbicki AT, Ream TS, Haag JR, Pikaard CS. 2009. RNA polymerase V transcription guides ARGONAUTE4 to chromatin. *Nat Genet.* 41: 630–634.
Wilson RC, Doudna JA. 2013. Molecular mechanisms of RNA interference. *Annu Rev Biophys.* 42:217–239.
Xie Z, Zilberman D, Johansen LK, Gustafson AM, Kasschau KD, Lellis AD, Jacobsen SE, Carrington JC. 2004. Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol.* 2:e104.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.
Zhang J, Wu T, Li L, Han S, Li X, Zhang S, Qi L. 2013. Dynamic expression of small RNA populations in larch (*Larix leptolepis*). *Planta* 237: 89–101.
Zhang X, Henderson IR, Lu C, Green PJ, Jacobsen SE. 2007. Role of RNA polymerase IV in plant small RNA metabolism. *Proc Natl Acad Sci U S A.* 104:4536–4541.